

Full Open Population Capture-Recapture Models with Individual Covariates

Matthew R. Schofield^{*†} and Richard J. Barker[‡]

Abstract

Traditional analyses of capture-recapture data are based on likelihood functions that explicitly integrate out all missing data. We use a complete data likelihood (CDL) to show how a wide range of capture-recapture models can be easily fitted using readily available software JAGS/BUGS even when there are individual-specific time-varying covariates. The models we describe extend those that condition on first capture to include abundance parameters, or parameters related to abundance, such as population size, birth rates or lifetime. The use of a CDL means that any missing data, including uncertain individual covariates, can be included in models without the need for customized likelihood functions. This approach also facilitates modeling processes of demographic interest rather than the complexities caused by non-ignorable missing data. We illustrate using two examples, (i) open population modeling in the presence of a censored time-varying individual covariate in a full robust-design, and (ii) full open population multi-state modeling in the presence of a partially observed categorical variable.

^{*}Department of Statistics, Columbia University, New York, NY, USA

[†]Current Address: Department of Statistics, University of Kentucky, Lexington, KY 40506, USA. E-mail: matthew.schofield@uky.edu

[‡]Department of Mathematics and Statistics, University of Otago, PO Box 56, Dunedin, New Zealand. E-mail: rbarker@maths.otago.ac.nz

Keywords: capture-recapture, demographic parameters, hierarchical modeling, individual covariates, JAGS/BUGS

1 Introduction

An important feature of capture-recapture modeling is the ability to include covariates, including individual-specific ones (Lebreton et al. 1992, Schwarz et al. 1993, Bonner and Schwarz 2006, King et al. 2008, Catchpole et al. 2008, Bonner et al. 2010). The development of methods for including individual covariates has focused on models that condition on the first capture of each individual. A consequence is that likelihood based inference is restricted to statements about survival or recapture probabilities and related quantities. Importantly, these models do not include abundance parameters (or parameters related to abundance, such as population growth rates or stopover time) in the likelihood. Instead, inference about abundance has relied on ad-hoc Horvitz-Thompson-type approaches (Huggins 1989, McDonald and Amstrup 2001).

Individual-specific time-varying covariates pose problems in models based on the full likelihood as the covariate values cannot be observed for individuals that were available for capture but were not caught. In order to fit these models by maximum likelihood, we would need to integrate out the missing covariate values for the unseen individuals for each possible value of abundance, as well as integrating out the missing values for the individuals we observed. Explicit integration of the likelihood leads to the observed data likelihood (ODL), but this integration is often difficult to do in practice. An alternative approach is to model in terms of the complete data likelihood (CDL), where specialized computational algorithms, such as the Gibbs sampler or the EM algorithm perform the required integration during the model fitting process. Schofield (2007), Schofield and Barker (2008, 2009) lay out a unified framework for capture-recapture modeling using the CDL,

which while flexible enough to include individual-specific time-varying covariates, still requires the user to construct these algorithms in order to fit the model. The same process is required when including individual covariates in closed population studies, for example, see King and Brooks (2008), Royle (2009).

To date, no user-friendly software is available for full open population modeling in the presence of individual-specific time-varying continuous covariates. Algorithms do exist for other individual-specific open population models, including user-written algorithms for the multi-state model (Dupuis and Schwarz 2007) and models where survival and fecundity parameters depend on population abundance (Schofield and Barker 2008). Unfortunately, these algorithms require custom-written software and lack generalizability. Another approach is to include these models in JAGS (Plummer 2003) or BUGS (Lunn et al. 2000), software that fits models using the Gibbs sampler and is being increasingly used by ecologists Royle et al. (2007), Gimenez et al. (2009), Schofield et al. (2009), Link and Barker (2010). This has been done for individual-specific mixed effects models in Royle and Dorazio (2008), Link and Barker (2010).

Here we consider models for two datasets, both of which have individual-specific time-varying covariates. The first is an example from Nichols et al. (1992), where the robust design is used to sample meadow voles, *Microtus pennsylvanicus*, with the body mass measured every time there was a capture. The body mass measurements were discretized in order to fit a multi-state model, conditional on first capture, to understand how the covariate body mass changed through time. This was later refined by Bonner and Schwarz (2006) who model body mass as an individual-specific time-varying covariate, although, their analysis ignored the robust design and conditioned on first capture. Other analyses have used the robust design to estimate abundance, but did this without taking into account body mass (Williams et al. 2002, Pg. 525). Here we use information on the body mass within the robust design to estimate abundance.

The second dataset is a study of conjunctivitis in house finch, *Carpodacus mexicanus*, where the covariate of interest is disease status, a covariate that is not always observed. Conn and Cooch (2009) model the data using an extension to the multi-state model to account for the partially observed data, however, they only model conditional on first capture. Even though they are able to obtain estimates of survival for each disease class and movement between the two disease classes, they were unable to get an predictions of the population in each disease class through time.

Here we show how the framework of Schofield (2007), Schofield and Barker (2008, 2009) can be used to extend the models of Royle et al. (2007), Royle and Dorazio (2008) and Link and Barker (2010) to fit full open population models with individual-specific time-varying covariates in BUGS. There are only minor differences in the model and BUGS code for the two examples, despite the differences between the datasets: (i) continuous vs. categorical covariate, (ii) robust design vs. standard capture-recapture design, (iii) covariate measurement every capture occasion vs. uncertainty in covariate measurements on some capture occasions. This shows the power and flexibility of the modeling approach, since the two examples include as special cases: the multi-state model, multi-event type models, individual-specific time-varying continuous covariates, and lifetime duration models such as the stop-over model.

2 Model Framework

Capture-recapture models involve complex missing data mechanisms. Traditional approaches to inference focus on deriving the likelihood for the observed data (the ODL) by integrating over all missing data. Instead, we use the modeling framework of Schofield (2007), Schofield and Barker (2008, 2009) that uses data augmentation (Tanner and Wong 1987) to allow us to model in terms of the complete data likelihood (CDL). Similar ideas

have also been proposed by Royle and Dorazio (2008).

The likelihood we use for inference is in terms of the complete data, which for a capture-recapture study with individual-specific time-varying covariate data are the (i) times of birth, (ii) times of death, and (iii) complete covariate values for each individual ever available for capture. The main advantage of using this likelihood over the ODL is that we are able to focus on modeling the processes of interest rather than having to account for the complexities caused by missing data that result from sampling methods. Importantly, in adopting the CDL approach to inference, we do not need to make any additional assumptions to those made when using the ODL. It is simply a reformulation of the model in terms of the easier-to-understand CDL where we use computational algorithms, such as Markov chain Monte Carlo (MCMC) or the expectation-maximization (EM) algorithm, to integrate over all missing data.

We write the CDL for the capture-recapture model including birth and covariates as

$$\underbrace{[a^b|\theta^b, N]}_{\text{Birth}} \underbrace{[a^d|a^b, \theta^d, N]}_{\text{Mortality}} \underbrace{[X|a^b, a^d, \theta^X, N]}_{\text{Capture}} \underbrace{[z|\theta^z, N]}_{\text{Covariate}}, \quad (1)$$

where $[Y|X]$ is the probability (density) of the random variable Y given X ;

$a_{ij}^b = 0$ means that individual i has yet to be born at, or before, sampling period j , with

$$a_{ij}^b = 1 \text{ otherwise;}$$

$a_{ij}^d = 1$ means that individual i has yet to die at sampling period j , with $a_{ij}^d = 0$ otherwise;

$X_{ij} = 1$ means that individual i was caught in sampling period j , with $X_{ij} = 0$ otherwise;

z_{ij} is the covariate value for individual i in sampling period j ;

θ^b are parameters describing the birth process, θ^d are parameters describing the mortality process, θ^X are parameters describing the capture process, θ^z are parameters of the covariate distribution and N is the total number of individuals available for capture during the

study period (which is distinct from N_j , the number of individuals alive in the j th sampling occasion). The covariate z can be used to help model the birth, death and covariate processes, although we assume that this is specified through the models for θ^b , θ^d and θ^X . In order for inference to be valid, we must ensure that these models do not violate the laws of conditional probability, for example, by assuming that survival probability depends on the covariate value at the end of the period.

Most demographic summaries of interest are obtained as derived quantities of a^d and a^b . For example, the population size in each sampling period, N_j , and the “lifetime” for each individual Δ_i are

$$N_j = \sum_{i=1}^N a_{ij}^b a_{ij}^d, \quad j = 1, \dots, k,$$

$$\Delta_i = \sum_{j=1}^k a_{ij}^b a_{ij}^d, \quad i = 1, \dots, N.$$

Other potential quantities of interest we could specify include the number of births and deaths between each sampling period.

The choice of model for each of the components in (1) will depend on the data, and the assumptions we are willing to make. For a standard capture-recapture study design, we present some common models for each component. We leave the parameters for mortality and capture to be individual specific as these are being modeled in terms of individual-specific covariates.

2.1 The Birth Component

One possible model for the birth components is

$$[a^b | \theta^b, N] = \prod_{i=1}^N [a_{i1}^b | \theta^b] \prod_{j=2}^k [a_{ij}^b | a_{i1}^b, \dots, a_{ij-1}^b, \theta^b]$$

$$= \prod_{i=1}^N \text{Bern}(\zeta_1) \prod_{j=2}^k \text{Bern} \left(\prod_{h=1}^{j-1} (1 - a_{ih}^b) \zeta_j + \left(1 - \prod_{h=1}^{j-1} (1 - a_{ih}^b) \right) \right), \quad (2)$$

where $\text{Bern}(p)$ denotes a Bernoulli distribution with parameter p and we set $\zeta_k = 1$. We include the term $\prod_{h=1}^{j-1} (1 - a_{ih}^b) \zeta_j + 1 - \prod_{h=1}^{j-1} (1 - a_{ih}^b)$ to ensure that an individual can only be born once. The value ζ_1 is the probability of being born before the study began, with the values ζ_{j+1} defined as the probability of being born between sample j and $j + 1$ conditional on (i) not being born before j , and (ii) being at risk of capture at some point during the study. A possible reparameterization is

$$\beta_0 = \zeta_1, \quad \beta_j = \zeta_{j+1} \left(1 - \sum_{h=1}^{j-1} \beta_h \right), \quad j = 1, \dots, k-1.$$

This gives the birth formulation of Schwarz and Arnason (1996) where β_j is the probability of being born between sample j and $j + 1$ conditional on being at risk of capture at some point during the study. Neither ζ_j nor β_j are meaningful birth parameters, since they are defined in terms of the study/sampling process, as for example, a change in k changes the parameter values. A more natural parameterization is to use per-capita birth rates,

$$\eta_j = \frac{\beta_j N}{N_j}, \quad j = 1, \dots, k-1.$$

This gives the birth formulation used in Schofield and Barker (2008), where η_j is the expected number of births between sampling period j and $j + 1$ for each individual in the population at sampling period j conditional on N . These parameters are similar to the birth parameters, f_j , used by Pradel (1996), Link and Barker (2005); the difference is that the denominator they use is $E[N_j|N]$ instead of N_j .

2.2 The Mortality Component

We factor the component for mortality as,

$$\begin{aligned}
[a^d|a^b, \theta^d, N] &= \prod_{i=1}^N \prod_{j=2}^k [a_{ij}^d | a_{ij-1}^d, a_{ij-1}^b, \theta^d] \\
&= \prod_{i=1}^N \prod_{j=2}^k \text{Bern}(a_{ij-1}^d (a_{ij-1}^b S_{ij} + (1 - a_{ij-1}^b))),
\end{aligned} \tag{3}$$

where the parameter S_{ij} is the probability of individual i surviving between sampling period j and $j + 1$. The term $a_{ij-1}^d (a_{ij-1}^b S_{ij} + (1 - a_{ij-1}^b))$ is required to ensure that an individual can only (i) die after being born, and (ii) live once.

2.3 The Capture Component

We factor the component for capture as,

$$\begin{aligned}
[X|a^b, a^d, \theta^X, N] &= \prod_{i=1}^N \prod_{j=1}^k [X_{ij} | a_{ij}^b, a_{ij}^d, \theta^X] \\
&= \prod_{i=1}^N \prod_{j=1}^k \text{Bern}(a_{ij}^d a_{ij}^b p_{ij}),
\end{aligned} \tag{4}$$

where p_{ij} is the probability of capture for individual i in sampling period j . The term $a_{ij}^d a_{ij}^b$ is required to ensure that an individual is only available for capture while it is alive.

2.4 Additions Required For BUGS/JAGS

In order to make inference about the parameters in the model, we adopt a Bayesian approach and fit all examples using the software JAGS (Plummer 2003), with model, data, initial values and script files available at www.maths.otago.ac.nz/~rbarker/BUGS. We use JAGS for the following examples due to superior convergence in trial runs of the algorithms as compared to OpenBUGS. The modeling language of JAGS is nearly identical

to BUGS (Lunn et al. 2000) and is able to be called from R (Su and Yajima 2009). We describe the main differences between BUGS and JAGS and how to specify the data and the initial values in the supplementary materials.

Neither JAGS nor BUGS allow stochastic indices, such as N , as was used in Schofield and Barker (2008). Instead, we must use a computational trick to include N . The trick is given in Durban and Elston (2005) and involves specifying M , an upper bound for N . An alternate, yet mathematically equivalent approach is given in Royle et al. (2007). They also specify M but then reparameterize the model in terms of an incomplete indicator variable, w , instead of N . The value $w_i = 1$ means that individual i was at risk of capture during the study and $w_i = 0$ otherwise, with $\sum_{i=1}^M w_i = N$. Having used both approaches, we find them practically equivalent with the exception that the specification of Royle et al. (2007) no longer has N available for hierarchical modeling, but does generally run slightly faster than the approach of Durban and Elston (2005). Since we have no desire to include a hierarchical model for N in the examples we explore, we use the approach of Royle et al. (2007) to make use of the faster algorithm. For readability, we do not change the CDL in (1) to reflect this additional likelihood component and continue to write the CDL conditioning on N .

Another difficulty is that attempting to use the per-capita birth rate (η_j) parameterization in JAGS or BUGS results in code that is currently impractically slow to run. Here we use the less natural ζ_j parameterization described above. This is not a limitation of MCMC or the Gibbs sampler, as Schofield and Barker (2008) used the per-capita birth rates. Instead it is a current limitation of JAGS and BUGS.

3 Example: Meadow Voles

Nichols et al. (1992) report a capture-recapture study of meadow voles, *Microtus pennsylv-*

vanicus, using a robust design (Pollock 1982) with 173 individuals caught over 6 primary periods each with 5 secondary samples. Every time an individual was caught the mass of the animal was recorded to the nearest integer. The scales used to measure mass had a maximum at 60 grams, with many individuals censored.

Bonner and Schwarz (2006) collapsed the data across the secondary periods, and allocated a single measurement to the observed body mass. Using these data they extended the Cormack-Jolly-Seber model (CJS) to include an individual-specific time-varying continuous covariate. Schofield et al. (2009) used the same data to show how to fit this model using BUGS (neither Bonner and Schwarz (2006) nor Schofield et al. (2009) accounted for the censored data). Here we extend this model to include the full robust design and make use of all information on body mass. Including the birth process in the model allows us to estimate the population size of the meadow vole in the presence of the individual-specific time-varying continuous covariate.

The CDL is given in (1), with the birth component given in (2) and the mortality component given in (3). To include the robust design, we redefine X to be an array with $X_{ijl} = 1$ if individual i is caught in primary period j and secondary period l . We extend the capture component in (4) to include both the k_1 primary periods and the k_{2j} secondary periods,

$$\begin{aligned} [X|a^b, a^d, \theta^X, N] &= \prod_{i=1}^N \prod_{j=1}^{k_1} \prod_{l=1}^{k_{2j}} [X_{ijl}|a_{ij}^b, a_{ij}^d, \theta^X] \\ &= \prod_{i=1}^N \prod_{j=1}^{k_1} \prod_{l=1}^{k_{2j}} \text{Bern}(a_{ij}^d a_{ij}^b p_{ijl}), \end{aligned}$$

We also redefine z to be an array with z_{ijl} being the body mass for individual i in primary period j and secondary period l . Since every observed body mass value was either censored or rounded we treat z_{ijl} as missing for every individual in every sampling period

and denote the observed masses by z_{ijl}^{obs} . We specify the model for z as

$$\begin{aligned} [z|\theta^z, N] &= \prod_{i=1}^N \prod_{j=b_i}^{k_1} \prod_{l=1}^{k_{2j}} [z_{ijl}|\theta^z] \\ &= \prod_{i=1}^N \prod_{j=f_i}^{k_1} \prod_{l=1}^{k_{2j}} N(\lambda_{ij}, \sigma_z^2) I(\delta_{1ijl}, \delta_{2ijl}), \end{aligned}$$

where $N(\mu, \sigma^2)$ denotes a Normal distribution with mean μ and variance σ^2 , b_i is the first sample individual i was alive and $I()$ is used to include the rounding, censoring and truncation, with

$$\delta_{1ijl} = \begin{cases} z_{ijl}^{obs} - 0.5 & \text{if } X_{ijl} = 1 \\ 0 & \text{otherwise} \end{cases}, \quad \delta_{2ijl} = \begin{cases} z_{ijl}^{obs} + 0.5 & \text{if } X_{ijl} = 1 \text{ and } z_{ijl}^{obs} \neq 60 \\ \infty & \text{otherwise.} \end{cases}$$

In other words, we assume that during the secondary periods when the population is assumed to be closed, mass remains constant and any differences are attributable to the measurement error σ_z^2 . We follow Bonner and Schwarz (2006) and model λ_{ij} as

$$\lambda_{ib_i} \sim N(\mu_\lambda, \sigma_{\lambda 1}^2), \quad \lambda_{ij} \sim N(\lambda_{ij-1} + \Delta_{j-1}, \sigma_{\lambda 2}^2), \quad j = b_i + 1, \dots, k_1.$$

This is a random walk with drift, where the mass of each individual increases, on average, Δ_j grams between primary period j and $j + 1$.

We model parameters S and p as

$$\begin{aligned} \text{logit}(S_{ij}) &= \alpha_0 + \alpha_1 \lambda'_{ij} + \eta_j^S, \quad \eta_j^S \sim N(0, \sigma_S^2), \\ \text{logit}(p_{ijl}) &= \gamma_0 + \gamma_1 \lambda'_{ij} + \eta_j^p + \epsilon_{jl}^p, \quad \eta_j^p \sim N(0, \sigma_{p1}^2), \quad \epsilon_{jl}^p \sim N(0, \sigma_{p2}^2), \end{aligned}$$

where λ'_{ij} is an approximately standardized value of λ_{ij} . We allow individuals to have survival probabilities that depend on their mass, with additional temporal variability, modeled

as a random effect. We allow probability of capture to depend on body mass due to allow for a sampling strategy that discriminated due to body mass, with additional variability within the secondary period and between primary periods, both modeled as a random effect. This is equivalent to specifying the closed population model M_t as a random effect, with the mean varying between the primary periods. The priors for all parameters are given in the supplementary materials.

In order to determine the effect that rounding and censoring have on the results we also fit the model with $z_{ijl} = z_{ijl}^{obs}$ when $X_{ijl} = 1$.

Each of the models was run in JAGS with an adaptive phase of 5000 iterations followed by a posterior sample of 20000 iterations. To ensure convergence we run 3 parallel chains with different starting values and checked convergence with the Brooks-Gelman-Rubin diagnostic (Brooks and Gelman 1998). We combined the posterior samples from the three chains to give a total posterior sample of 60000.

The results suggest that there is little practical difference between accounting for the censoring of the covariate values and simply modeling using the raw observations (figure 1). While there are differences in the model for mass, particularly in the variances, this does not translate to substantial differences in the model for survival or probability of capture. While mass appeared to be associated with the probability of capture, with larger animals having a higher chance of capture, there is no evidence of body mass being associated with survival. This result differs from Bonner and Schwarz (2006) and Schofield et al. (2009) who found that body mass was not associated with either capture probability or survival. This difference appears to be due to Bonner and Schwarz (2006) compressing the robust design into a more standard CJS design. To ensure that this is consistent with the data we compared the observed data for individuals caught at least once in any given secondary period, with a significant increase in the average body mass as the number of captures increases. After adjusting for the effect of body mass on capture probability, the abundance

of the meadow vole appears stable, fluctuating between ~ 55 to ~ 85 individuals during the study (figure 1).

4 Example: Conjunctivitis in House Finch

Conn and Cooch (2009) used a multi-state model to study conjunctivitis in house finch, *Carpodacus mexicanus*, with 813 individuals caught in 16 samples. A two-state model (whether or not an individual had conjunctivitis) was used, with some individuals having unknown status. Our approach is to include all missing disease information using data augmentation and treating the disease as a individual-specific time-varying categorical covariate (Dupuis 1995). Thus we can use the CDL in (1) with disease being the covariate z . Since the covariate takes two values, we can examine abundance, or any other demographic summary, separately for each group.

Dupuis and Schwarz (2007) used the CDL to fit a multi-state model and estimate abundance. Their approach differs to ours due to different computational algorithms. To improve the efficiency of their MCMC sampling, they summed over the latent state variables z_{ij} when updating N . While this will yield a quicker algorithm, both in terms of mixing and time, it lacks generalizability beyond the multi-state model to, say, individual-specific time-vary continuous covariates. In contrast, our approach allows us to apply the CDL across a range of different models and different covariate distributions, including continuous covariates, without major modifications in the JAGS/BUGS code.

Since many individuals have uncertain disease status, even when encountered, we must consider assumptions about the missingness of these observations (Rubin 1976). Here we assume that they are either missing completely at random or missing at random. In either case, the process that describes how the data go missing does not need to be included in the model. In the supplementary materials we describe and fit the model where we assume

that the additional missing data is missing not at random. The results are very similar to those from the model assuming the additional missing data is missing at random.

The CDL is given in (1) and we specify the birth component as in (2), the death component as in (3) and the capture component as in (4). The only component left to specify is the covariate, disease.

We specify the model for disease as

$$\begin{aligned} [z|\theta^z, N] &= \prod_{i=1}^N [z_{ib_i}|\theta^z] \prod_{j=b_i+1}^k [z_{ij}|z_{ij-1}, \theta^z] \\ &= \prod_{i=1}^N \text{Cat}((\nu_1, \nu_2)) \prod_{j=b_i+1}^k \text{Cat}((\omega_{z_{ij-1}1}, \omega_{z_{ij-1}2})) \end{aligned}$$

where $\text{Cat}(\boldsymbol{\pi})$ is a categorical distribution with probability vector $\boldsymbol{\pi}$. The covariate value $z_{ij} = 2$ indicates that individual i does has the disease in sample j and $z_{ij} = 1$ otherwise, ν_l is the probability of being in state l in the first sample after birth and ω_{hl} is the probability of moving from state h to state l . We use a generic categorical distribution (instead of the binomial) to show how this model generalizes to more than two states.

To complete the model specification we model S and p as

$$\begin{aligned} \text{logit}(S_{ij}) &= \alpha_0 + \alpha_1 I(z_{ij} = 2) + \eta_j^S, \quad \eta_j^S \sim N(0, \sigma_S^2), \\ \text{logit}(p_{ij}) &= \gamma_0 + \gamma_1 I(z_{ij} = 2) + \eta_j^p, \quad \eta_j^p \sim N(0, \sigma_p^2). \end{aligned}$$

We allow individuals to have survival and capture probabilities that depend on their disease status, with temporal variability modeled as a random effect. The priors for all parameters are given in the supplementary materials.

Each of the models was run in JAGS with an adaptive phase of 25000 iterations followed by a posterior sample of 25000 iterations. To ensure convergence we run 3 parallel chains with different starting values and checked convergence with the Brooks-Gelman-Rubin

diagnostic (Brooks and Gelman 1998). We combined the posterior samples from the three chains to give a total posterior sample of 75000.

The results suggest that disease status is associated with both survival and capture probability (figure 2). Having conjunctivitis lowered the log-odds of survival, while increasing the log-odds of capture. The results for survival appear to agree with Conn and Cooch (2009), with no results for capture probability available to compare. The transition probabilities suggest that it is rare for an individual to develop conjunctivitis, with less than 5% of disease free animals contracting the disease (figure 2). However, once an individual has conjunctivitis it is somewhat difficult to become disease free, with around three quarters of individuals remaining in the disease state from one sample to the next. The population size for diseased animals, while low, appears to be relatively stable with no fewer than 5 diseased individuals in the population (and as many as 40) during the study (figure 2). It is interesting to note that the population sizes for the diseased and non-diseased states, while similar, do not necessarily exhibit the same dynamics through time. In particular, one could claim that there are periods where one class increases or decreases while the other remains relatively stable.

5 Discussion

We have shown how to use the framework of Schofield (2007), Schofield and Barker (2008, 2009) to fit using JAGS or BUGS complex models where we estimate demographic summaries of interest in the presence of individual-specific time-varying covariates. The two examples we show include modeling in the presence of covariate uncertainty and including different study designs. The CDL conveniently factorizes so that we are able to specify separate models for the birth, mortality, capture and covariate processes while fitting the joint model in JAGS or BUGS using MCMC methods. This means our focus can move from

accounting for the complex sampling process to focusing on specifying biologically meaningful models for the processes of interest, including hierarchical models for the parameters that describe demographic changes in the population.

The house finch example shows how using the CDL facilitates modeling of missing data, including covariate uncertainty. The model is identical to the one where all data were actually observed, except that we must now consider, and potentially include, the process by which the data go missing. This is in contrast to the ODL where any missing data, including covariate uncertainty, usually require specification of a new likelihood. Examples of this can be seen in Nichols et al. (2004) and Conn and Cooch (2009) where uncertainty in the covariate requires a new likelihood to be specified in order to include the missing data.

The advantages in using the CDL for modeling comes at a computational cost. In a Bayesian setting, all missing values are treated as ‘unknowns’, which means each missing value needs to be updated in every MCMC iteration. Thus the computational burden increases with the amount of missing data. With current processor power, we are limited in the size of the data sets that we are able to fit. Using JAGS/BUGS, it can soon become difficult to fit datasets with either a large number of individuals or many sampling periods. To a large extent, these deficiencies can be overcome with user-written, application specific code, making use of specialized algorithms and other computational advantages (such as Dupuis and Schwarz (2007) for the multi-state model). However, this limits generalizability, with new algorithms needed for each application. These computational limitations, while serious, should not be an impediment to use. Advances in algorithms for MCMC and continuing increases in computational power mean that we will continue to be able to fit bigger and more complex models into the future.

An issue we have not mentioned is model checking. We suggest that model checking be done using posterior predictive checking (Gelman et al. 2004, Pg. 159). One approach

is to focus on features of the data that are biologically important. For example, we may want to ensure that our model explains the difference between the number of individuals caught in successive sample occasions. We then generate replicate datasets from the fitted model and either (i) visually check, or (ii) specify an appropriate test-statistic, to see if the replicates are consistent with the observed data. An example of a visual check is shown in (Gelman et al. 2004, Pg. 164-165) with the speed of light data.

An alternative approach is to make use of an omnibus test statistic to ensure that our fitted model does not generate data that are inconsistent with the data we have observed. Approaches that have been adopted in the Bayesian mark-recapture literature include use of the likelihood function (King and Brooks 2002) or Freeman-Tukey statistic (Brooks et al. 2000).

Goodness-of-fit for mark-recapture models has received considerable attention in a frequentist setting (Pollock et al. 1985, Burnham et al. 1987, Barker 1999, Pradel et al. 2003). A constructive approach to goodness-of-fit testing can be taken when a multinomial model is used that is a member of the full exponential family based on the factorization

$$[\text{Data} \mid \text{MSS}] \times [\text{MSS} \mid \boldsymbol{\theta}].$$

For a large class of mark-recapture models the term $[\text{Data} \mid \text{MSS}]$ has a hypergeometric distribution and provides a natural partitioning of the data into test components. As far as we are aware an approach based on the predictive distribution $[\text{Data} \mid \text{MSS}]$ has been little used in a Bayesian setting (see Wright et al. 2009, for an exception) but we believe that this should be a productive approach to goodness-of-fit assessment.

A related issue is model selection. We recommend different approaches to model selection depending on the objective of the study. If the objective is to learn about the system and to generate hypotheses about relationships then we advocate exploring the data and

finding models that best fit the data. A number of models can be explored and compared using cross-validation (Hastie et al. 2009) as well as other criteria. If the objective is in validating previously generated hypotheses and making inference in the presence of model uncertainty, then we advocate exploring posterior model probabilities, or equivalently Bayes Factors, for a small set of scientifically driven models (Link and Barker 2006). Barker and Link (2010) outline an approach to calculating posterior model probabilities that uses the output obtained from running the individual models using MCMC. Other techniques for calculating posterior model probabilities are described in King et al. (2010). Another possible approach to model selection involves specifying a hierarchical model, that has as special cases, all models considered (Gelman et al. 2004, Pg. 405 – 406). Then, instead of the usual approach of either including the effect or not including the effect before potentially combining the results using model-averaging, we can specify an informative prior distribution centered on zero, that can be viewed as a compromise between inclusion of the effect (with an approximately flat prior) and exclusion of the effect (with a prior with all mass at zero). An example of this approach is when we have a potential time effect. Instead of choosing between a model with no time effect, and one with a separate and unrelated parameter for each time point, we can, as we do in the two examples here, include time as a random effect with the variance estimated from the data.

The CDL that we used to fit all models here can be extended in a number of ways. We expect it to extend naturally to continuous data models. Depending on the observed data, we would expect the CDL to remain the same, or at least similar, with the conditional likelihood components relating to birth, death, capture and any covariates changing to account for continuous time processes.

Another extension is when we have uncertainty in the tags themselves. An example of this is when our “tag” is a DNA profile of the individual. The problem here is that uncertainty in the DNA profile is due to various genotyping errors. The only change we

require in our CDL is to include the true tag as missing data and include a component that describes the corruption of the true tags to the observed tags. Wright et al. (2009) used this approach to estimate population size in a closed population study.

Using the CDL, we are able to model complex dependencies between individuals in the population. For example, Schofield and Barker (2008) included density dependence on both birth rates and survival probabilities. Another potential example is one where DNA information for one individual is able to provide information about another individual, such as parents and offspring. Information about the death of parent gives information about the time of birth of the offspring and vice versa. The CDL approach is, at least in principle, able to include these relationships as well as including potential uncertainty in the offspring/parent relationship.

References

- Barker, R. (1999), “Joint analysis of mark-recapture, resighting and ring-recovery data with age-dependence and marking-effect,” *Bird Study*, 46(suppl), S82–91.
- Barker, R. J. and Link, W. A. (2010), “Posterior model probabilities computed from model-specific Gibbs output,” In Preparation.
- Bonner, S. J., Morgan, B. J. T., and King, R. (2010), “Continuous Covariates in Mark-Recapture-Recovery Analysis: A Comparison of Methods,” *Biometrics*, In Press.
- Bonner, S. J. and Schwarz, C. J. (2006), “An Extension of the Cormack-Jolly-Seber Model for Continuous Covariates with Application to *Microtus pennsylvanicus*,” *Biometrics*, 62, 142–149.
- Brooks, S. P., Catchpole, E. A., and Morgan, B. J. T. (2000), “Bayesian Animal Survival Estimation,” *Statistical Science*, 15, 357–376.

- Brooks, S. P. and Gelman, A. (1998), “General Methods for Monitoring Convergence of Iterative Simulations,” *Journal of Computational and Graphical Statistics*, 7, 434–455.
- Burnham, K. P., Anderson, D. R., White, G. C., Brownie., and Pollock, . K. (1987), “Design and analysis methods for fish survival experiments based on release-recapture.” *American Fisheries Society Monograph*, 5, 1–437.
- Catchpole, E., Morgan, B., and Tavecchia, G. (2008), “A new method for analysing discrete life history data with missing covariate values,” *Journal of the Royal Statistical Society: Series B(Statistical Methodology)*, 70, 445–460.
- Conn, P. and Cooch, E. (2009), “Multistate capture-recapture analysis under imperfect state observation: an application to disease models,” *Journal of Applied Ecology*, 46, 486–492.
- Dupuis, J. and Schwarz, C. (2007), “A Bayesian approach to the multistate Jolly-Seber capture-recapture model,” *Biometrics*, 63, 1015–1022.
- Dupuis, J. A. (1995), “Bayesian estimation of movement and survival probabilities from capture-recapture data,” *Biometrika*, 82, 761–772.
- Durban, J. W. and Elston, D. A. (2005), “Mark-Recapture With Occasion and Individual Effects: Abundance Estimation Through Bayesian Model Selection in a Fixed Dimensional Parameter Space,” *Journal of Agricultural, Biological, and Environmental Statistics*, 10, 291–305.
- Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004), *Bayesian Data Analysis*, Chapman & Hall/CRC, 2nd ed.
- Gimenez, O., Bonner, S. J., King, R., Parker, R. A., Brooks, S. P., Jamieson, L. E., Grosbois, V., Morgan, B. J. T., and Thomas, L. (2009), “WinBUGS for population

- ecologists: Bayesian modeling using Markov Chain Monte Carlo methods,” in *Modeling Demographic Processes in Marked Populations*, eds. Thomson, D. L., Cooch, E. G., and Conroy, M. J., Springer, vol. 3 of *Environmental and Ecological Statistics*, pp. 885–918.
- Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The Elements of Statistical Learning. Second Edition.*, Springer.
- Huggins, R. (1989), “On the Statistical Analysis of Capture Experiments,” *Biometrika*, 76, 133 – 140.
- King, R., Brooks, S., and Coulson, T. (2008), “Analyzing Complex Capture–Recapture Data in the Presence of Individual and Temporal Covariates and Model Uncertainty,” *Biometrics*, 64, 1187–1195.
- King, R. and Brooks, S. P. (2002), “Bayesian model discrimination for multiple strata capture-recapture data,” *Biometrika*, 89, 785–806.
- (2008), “On the Bayesian Estimation of a Closed Population Size in the Presence of Heterogeneity and Model Uncertainty,” *Biometrics*, 64, 816–824.
- King, R., Morgan, B. J. T., Gimenez, O., and Brooks, S. P. (2010), *Bayesian Analysis for Population Ecology*, Chapman & Hall/CRC.
- Lebreton, J. D., Burnham, K., Clobert, J., and Anderson, D. (1992), “Modelling survival and testing biological hypotheses using marked animals: A unified approach with case studies,” *Ecological Monographs*, 62, 67–118.
- Link, W. A. and Barker, R. J. (2005), “Modeling Association among Demographic Parameters in Analysis of Open Population Capture-Recapture Data,” *Biometrics*, 61, 46 – 54.

- (2006), “Model Weights and the Foundations of Multimodel Inference,” *Ecology*, 87, 2626–2635.
- (2010), *Bayesian Inference with ecological applications*, Academic Press.
- Lunn, D., Thomas, A., Best, N., and Spiegelhalter, D. (2000), “WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility,” *Statistics and Computing*, 10, 325–337.
- McDonald, T. and Amstrup, S. (2001), “Estimation of population size using open capture-recapture models,” *Journal of Agricultural, Biological, and Environmental Statistics*, 206–220.
- Nichols, J., Sauer, J., Pollock, K., and Hestbeck, J. (1992), “Estimating transition probabilities for stage-based population projection matrices using capture-recapture data,” *Ecology*, 73, 306–312.
- Nichols, J. D., Kendall, W. L., Hines, J. E., and Spendelov, J. A. (2004), “Estimation of Sex-Specific Survival from Capture-Recapture Data when Sex is not Always Known,” *Ecology*, 85, 3192–3201.
- Plummer, M. (2003), “JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling,” in *Proceedings of the 3rd International Workshop on Distributed Statistical Computing, March*, pp. 20–22.
- Pollock, K. H. (1982), “A Capture-Recapture Design Robust to Unequal Probability of Capture,” *Journal of Wildlife Management*, 46, 752 – 757.
- Pollock, K. H., Hines, J. E., and Nichols, J. D. (1985), “Goodness-of-Fit Tests for Open Capture-Recapture Models,” *Biometrics*, 41, 399–410.

- Pradel, R. (1996), “Utilization of Capture-Mark-Recapture for the Study of Recruitment and Population Growth Rate,” *Biometrics*, 52, 703–709.
- Pradel, R., Wintrebert, C., and Gimenez, O. (2003), “A proposal for a goodness-of-fit test to the Arnason-Schwarz Multisite capture-recapture model.” *Biometrics*, 59, 43–53.
- Royle, J. (2009), “Analysis of capture-recapture models with individual covariates using data augmentation,” *Biometrics*, 65, 267–274.
- Royle, J. and Dorazio, R. (2008), *Hierarchical modeling and inference in ecology: the analysis of data from populations, metapopulations and communities*, Academic Press.
- Royle, J. A., Dorazio, R. M., and Link, W. A. (2007), “Analysis of multinomial models with unknown index using data augmentation,” *Journal of Computational and Graphical Statistics*, 16, 67–85.
- Rubin, D. B. (1976), “Inference and missing data,” *Biometrika*, 63, 581–592.
- Schofield, M. R. (2007), “Hierarchical Capture-Recapture Models,” Ph.D. thesis, University of Otago.
- Schofield, M. R. and Barker, R. J. (2008), “A unified capture-recapture framework,” *Journal of Agricultural, Biological and Environmental Statistics*, 13, 458–477.
- (2009), “A Further Step Toward the Mother-of-all-Models: Flexibility and Functionality in the Modeling of Capture-Recapture Data,” in *Modeling Demographic Processes in Marked Populations*, eds. Thomson, D. L., Cooch, E. G., and Conroy, M. J., Springer, vol. 3 of *Environmental and Ecological Statistics*, pp. 677–689.
- Schofield, M. R., Barker, R. J., and MacKenzie, D. I. (2009), “Flexible hierarchical mark-recapture modeling for open populations using WinBUGS,” *Environmental and Ecological Statistics*, 16, 369–387.

- Schwarz, C. J. and Arnason, A. N. (1996), “A General Methodology for the Analysis of Capture-Recapture Experiments in Open Populations,” *Biometrics*, 52, 860–873.
- Schwarz, C. J., Schweigert, J. F., and Arnason, A. N. (1993), “Estimating migration rates using tag-recovery data,” *Biometrics*, 49, 177–193.
- Su, Y.-S. and Yajima, M. (2009), *R2jags: A package for running jags from R*, R package version 0.01-26.
- Tanner, M. A. and Wong, W. H. (1987), “The Calculation of Posterior Distributions by Data Augmentation (with discussion),” *Journal of the American Statistical Association*, 82, 529–550.
- Williams, B., Nichols, J., and Conroy, M. (2002), *Analysis and management of animal populations.*, Academic Press.
- Wright, J., Barker, R., Schofield, M., Frantz, A., Byrom, A., and Gleeson, D. (2009), “Incorporating genotype uncertainty into mark-recapture-type models for estimating abundance using DNA samples,” *Biometrics*, 65, 833–840.

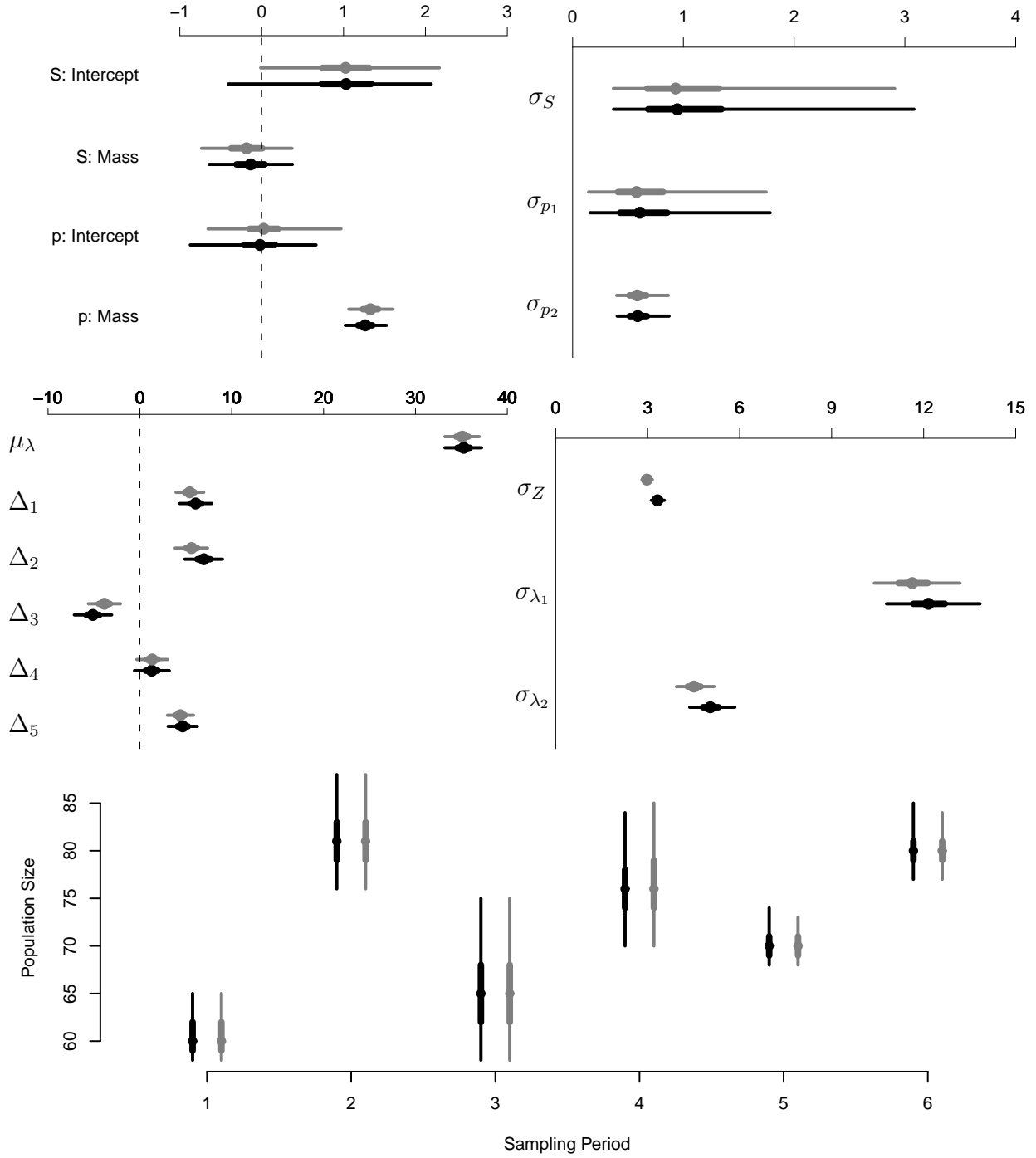


Figure 1: Parameter estimates for the meadow vole example. The point gives the median of the marginal posterior distribution and the lines represent the central 50% and 95% credible intervals. In all plots, black is the model with censoring and blue is the model without censoring.

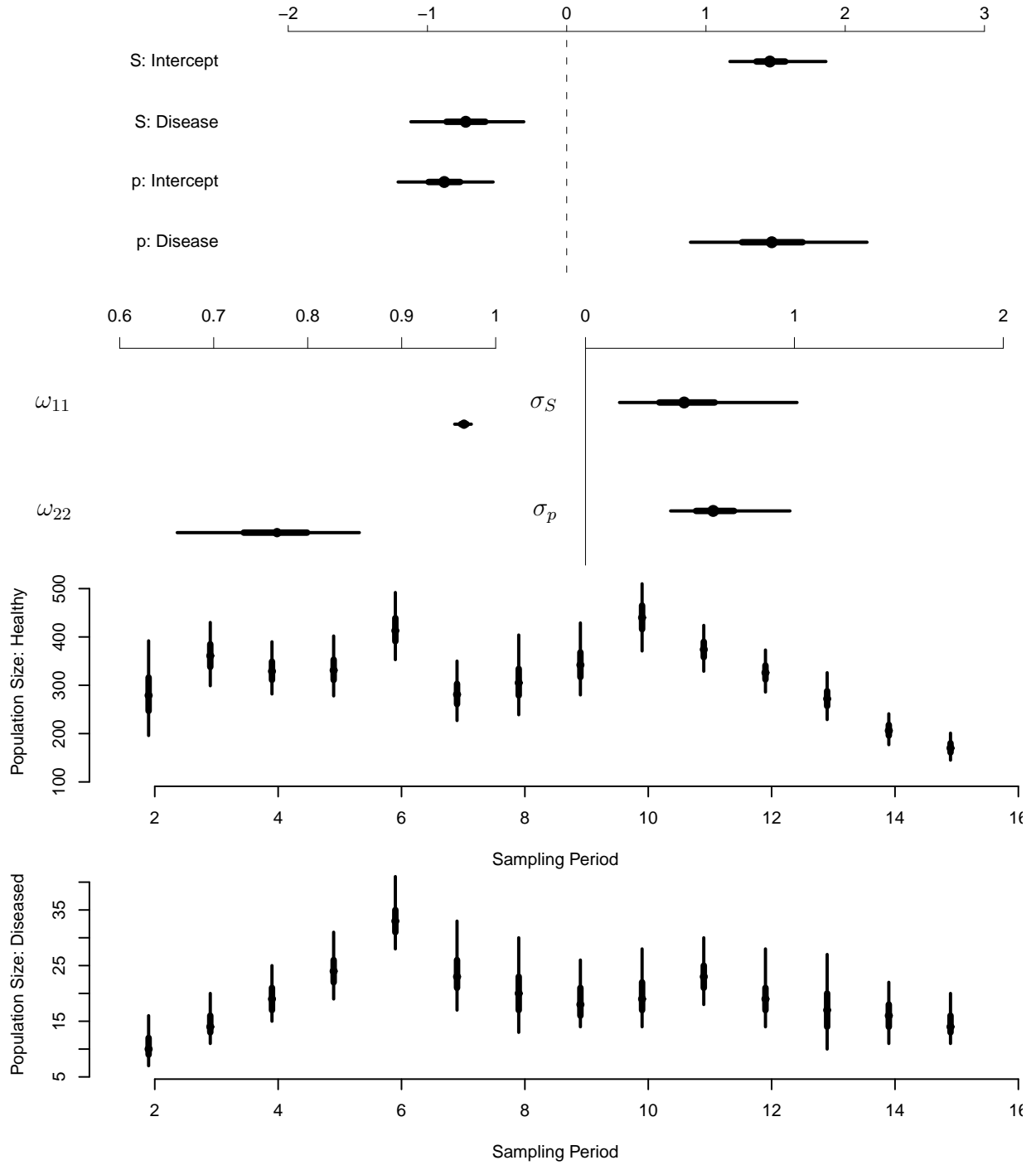


Figure 2: Parameter estimates for the house finch example. The point gives the median of the marginal posterior distribution and the lines represent the central 50% and 95% credible intervals.